Synopsis for Registration as Research Student

For

Doctoral Degree

On

"Design and Develop Stroke Prediction Using Imbalanced Classification."

By

Ms. Manisha Amrutrao Manjramkar

Research Scholar

Supervisor

Dr. Mrs. Kalpana Jondhale_Paithane
Professor

Department of Computer Science & Engineering

M. G. M's College of Engineering

Nanded-431605. (MS.) India.

E-mail: mboywar@gmail.com , Phone No.: 9404816610

Table of Contents

Abstract 1

1. Introduction 2

1.1 Stroke Causes 3

1.1.1 Ischemic Stroke 3

1.1.2 Hemorrhagic Stroke 4

1.2 Factors Associated with Stroke 4

1.2.1 Non- changeable Risk Factors 4

1.2.2 Changeable Risk Factors 4

1.3 Data Imbalance 6

2. Literature Review 7

3. Objectives of research 10

4. Methodology 11

4.1 Data Collection 11

4.2 Imbalanced Data Handling 12

4.3 Feature Selection and Engineering 12

4.4 Model Development 12

4.5 Evaluation Metrics 12

4.6 Integration, Validation and Testing 12

4.7 Proposed work plan 12

5. Ethical Considerations 14

References 15

Abstract

A stroke, also known as a cerebrovascular accident (CVA), is a medical disorder that happens, when the blood flow to the brain is stopped or diminished, causing brain cell damage or death [1]. It's a medical emergency that has to be addressed right away. Stroke is a life-threatening medical illness that need prompt medical intervention. Recognizing the signs and symptoms of stroke, knowing the risk factors, and implementing preventative actions are all critical in lowering the incidence and effect of stroke. Stroke risk factors include diabetes, high blood pressure, weight gain, elevated cholesterol, familial past with stroke, age (potent risk factors increases with age), and other health diseases [2]. A healthy lifestyle, risk factor management, and obtaining proper medical treatment can all help lower the incidence of stroke.

Stroke is a prominent cause of morbidity and death globally, and early identification and prediction of stroke cases are critical for prompt management and improved patient outcomes [3] [4]. However, the unbalanced nature of stroke prediction datasets, with a large proportion of "normal" cases and a small proportion of "stroke" cases, creates considerable hurdles in constructing reliable prediction models. The goal of this research proposal is to construct and develop a stroke prediction model employing imbalanced classification techniques in order to overcome the restrictions of existing methodologies.

The class imbalance problem in stroke prediction datasets affects the effectiveness of classical classification techniques. These models have a bias towards the majority class, ensuing in low stroke prediction accurateness. As a result, innovative imbalanced classification approaches particularly specialized for stroke prediction are required to enhance the diagnosis and early intervention of stroke patients. This proposed work plan outlines a comprehensive approach to address the imbalanced nature of stroke prediction with machine learning and data analysis methods.

The following are the goals of this research project:

- Create a unique imbalanced classification system for stroke prediction.
- Test the proposed algorithm's performance against a large and representative stroke prediction dataset.

- Compare the proposed algorithm's findings with current stroke prediction algorithms [5].
- Determine the clinical relevance and practical usefulness of the generated stroke prediction model.

This proposed work plan outlines a systematic approach to tackle the challenge of stroke prediction in imbalanced datasets. By leveraging advanced machine learning techniques, data preprocessing, and ethical considerations, this research aims to provide a valuable tool for early stroke risk assessment and prevention in clinical practice.

1. Introduction

Stroke, often referred to as a "brain attack," is a health emergency that arises as soon as there is an unexpected interruption of blood flow to the brain [6], resulting in damage to brain cells [1]. It is a leading cause of death and disability worldwide, with significant implications for individuals, families, and healthcare systems. Understanding stroke is crucial, as timely intervention can make the dissimilarity between life and demise, as well as between complete recovery and permanent disability. Stroke is a global healthiness fear, worrying millions of people each year. It is a leading cause of demise in many nations and ranks among the top causes of long-term disability [8]. The consequences of stroke can be devastating, affecting not only the individual's health but also their quality of life and independence [3].

Several risk factors contribute to the development of stroke, including high blood pressure, smoking, diabetes, obesity, and a sedentary lifestyle [9]. Genetic factors and family history can also play a role. Recognizing the warning signs of stroke is critical for prompt medical attention. Unexpected numbness or weakness in the face, arm, or leg; unexpected disorientation, trouble saying things, or difficulty comprehending speech; unexpectedly trouble vision in a single or both eyes; sudden acute headache; and problems with walking, fainting, or difficulties with balance and coordination are also typical indications [5].

"Time is brain" is a phrase often associated with stroke. The sooner a person experiencing a stroke receives medical treatment, the better the chances of minimizing brain damage and achieving a positive outcome. Rapid assessment and treatment, including the administration of clot-busting medications when appropriate, can be lifesaving. Stroke survivors often face a lengthy rehabilitation

process to regain lost functions and adapt to new challenges. Rehabilitation may include physical therapy, occupational therapy, speech therapy, and psychological support [1] [11].

Stroke is not only a medical crisis but also a preventable condition. Adopting a balanced diet, remaining healthy, treating long-term illnesses, and staying away from tobacco products and excessive consumption of alcohol are all examples of lifestyle changes that can dramatically lower the probability of stroke.

Advances in medical technology, research, and treatment options have improved outcomes for stroke patients in recent years. These include telemedicine for remote stroke assessments, minimally invasive procedures, and innovative rehabilitation techniques [1].

In this introduction to stroke, we've outlined the critical aspects of this medical condition, emphasizing its prevalence, types, risk factors, warning signs, and the importance of rapid intervention and rehabilitation [12]. Stroke is a complex and multifaceted health challenge, but with increased awareness, prevention efforts, and advancements in healthcare, we can work towards reducing its impact on individuals and communities.

1.1 Stroke Causes

A stroke is caused by the interruption or restriction of blood flow to the brain, leading to brain cell damage and potential neurological deficits. The specific causes of strokes can be broadly categorized into two main types: ischemic stroke and hemorrhagic stroke [13].

1.1.1 Ischemic Stroke [14]

- Thrombotic Stroke: Such a kind of ischemic stroke happens when a clot of blood develops inside a single of the brain's blood vessels, generally as a result of deposits of fatty substances (atherosclerosis) accumulating in the arteries delivering bloodstream into the brain. A clot may restrict flow of blood, resulting in a stroke.
- Embolic Stroke: A clot of bleeding blood or other debris originates somewhere in the body (typically in the heart) and flows through the bloodstream until it gets stuck in a smaller brain artery, producing a blockage.

1.1.2 Hemorrhagic Stroke [15]

- Intracerebral Hemorrhage: This type of stroke occurs when a blood vessel within the brain ruptures, leading to bleeding within the brain tissue. The most common cause of intracerebral hemorrhage is high blood pressure (hypertension), which weakens blood vessel walls over time.
- Subarachnoid Hemorrhage: Subarachnoid hemorrhage involves bleeding into the space between the brain and the thin membrane covering the brain (the subarachnoid space). It is often caused by the rupture of an aneurysm (a weakened and ballooned blood vessel) in the brain.

1.2 Factors Associated with Stroke

Stroke is a complex medical disorder that is influenced by a variety of damaging factors, both changeable and non-changeable. Understanding these risk factors is essential for stroke prevention and early intervention [16]. The associated risk factors for stroke are divided into two categories: changeable and non-changeable.

1.2.1 Non- changeable Risk Factors

- Age: The risk of stroke increases with age. While stroke can occur at any age, the likelihood of experiencing a stroke rises significantly as individuals get older [17].
- Gender: Stroke affects both men and women, but the risk factors may differ by gender. Men are generally at a slightly higher risk of stroke, particularly at younger ages. However, women's risk increases after menopause.
- Family History: A family history of stroke or certain cardiovascular conditions can increase an individual's risk. Genetics can play a role in predisposing someone to stroke.
- Race and Ethnicity: Some ethnic groups, such as African Americans, Hispanic Americans, and Native Americans, have a higher risk of stroke. This may be due to genetic factors and a higher prevalence of other stroke risk factors in these populations.

1.2.2 Changeable Risk Factors

- High Blood Pressure (Hypertension): The one and only significant changeable risk factor for stroke is a high blood pressure level. Vessels that carry blood can be damaged by hypertension, resulting in them being more vulnerable to clots or ruptures.
- Smoking: Tobacco usage, particularly cigarette use and being exposed to smoke from others, significantly raises the likelihood of stroke. Tobacco use destroys blood arteries and increases the production of blood clots.
- Diabetes: Uncontrolled diabetes can contribute to atherosclerosis (narrowing and hardening of arteries) and increase the risk of clot formation, increasing the likelihood of stroke.
- High Cholesterol: Elevated low-density lipoprotein cholesterol levels can promote plaque accumulation in blood vessels, increasing the risk of stroke.
- Obesity: Having overweight is linked to a number of health risks for stroke, notably diabetes, hypertension, and elevated cholesterol.
- Physical Inactivity: A sedentary lifestyle contributes to various risk factors for stroke, including obesity and high blood pressure. Regular physical activity can help reduce the risk.
- Unhealthy Diet: Saturated fat and Trans fat-rich diets, sodium (salt), and low in fruits, vegetables, and whole grains can contribute to stroke risk by promoting the development of atherosclerosis and hypertension.
- Excessive Alcohol Consumption: Frequent alcohol use can elevate blood pressure and add to other risk factors for cardiovascular disease, increasing the risk of stroke.
- Atrial Fibrillation: This irregular heartbeat condition can lead to the formation of blood clots in the heart, which can travel to the brain and cause a stroke.
- Drug Abuse: The use of illicit drugs, particularly stimulants like cocaine or amphetamines, can increase the risk of stroke due to their impact on blood vessels and blood pressure.
- Sleep Apnea: Untreated sleep apnea, a condition characterized by interrupted breathing during sleep, can be associated with an increased risk of stroke.

- Migraines with Aura: Individuals who experience migraines with aura may have a slightly higher risk of ischemic stroke.

It's important to note that many of these modifiable risk factors can be managed through lifestyle changes, medications, and regular medical check-ups [6] [18]. Stroke prevention strategies often focus on controlling these factors to reduce the risk and improve overall cardiovascular health. Individuals concerned about their stroke risk should consult with healthcare professionals for personalized guidance and risk assessment.

1.3 Data Imbalance

Data is acquired in everyday life from many areas such as social media, online sites, recordings, and databases. When working with information from multiple sources, it is important to evaluate the data quality [19] [20]. Data of unclear quality might describe numerous forms of preconceptions at multiple angles of the data science process. These preconceptions can sometimes influence the interaction of variables and, throughout many situations, can reveal the inverse of the real behavior [2] [21]. Data imbalances are major reasons of inaccuracy in the assessment of outcomes. A difficulty that happens frequently with doing categorization [3] [22].

It should also be mentioned that most data deteriorate from the drawback of imbalance, which implies that one of a group has a greater proportion than another group. The term "unbalanced dataset" refers to a dataset with an uneven class partition [5] [23].

A common concern is the classification of two classes. In such cases, the overriding class is seen as a negative class or majority, whereas underprivileged class is regarded as positive class or minority. Such forms of datasets exacerbate retrieval of information and processing problems [24] [25] [26] resulting in poor data quality [27]. Many researchers investigated comparable class imbalance (CI) applications, deceptive telephone conversations [28], text classification [29] [30], telecommunications management [28] and identification of spillage of oil in pictures of satellite [31] [32]. The classification machine learning or data mining techniques are based on two assumptions: the first is that output accuracy should be maximized, and the second is that test data should be selected from the distribution of the training data [33]. When there is an imbalance in the data, one or both assumptions are broken

[34]. To further appreciate the issue of CI, consider the example of fraud detection [28].

Assume that based on the data supplied, a classifier system forecasts fraudulent transactions. Only a small number of individuals engage in fraudulent transactions in everyday life. And the majority of individuals in the records will have non-fraudulent transactions. Due to the uneven proportion of the classification process in the data, the suggested classifier will always regard a large number of fraudulent transactions as non-fraudulent transactions. The classifier system will generate a large number of correct answers but decreased productivity. As a result, it is vital to tackle this problem and establish correct mechanisms before concentrating on categorization technique selection. The goal of this study is to learn and apply typical techniques to solving this problem.

The foregoing discussion defined minority and majority classes and the requirement to handle unbalanced data. The emphasis in this part will be on the variables that make studying unbalanced dataset challenging. The imbalance issue, according to Japkowicz et al. [35], is guided by four elements i.e. the degree of CI, the difficulty in properly understanding the data, the total number of training sample, and the kind of classification technique [36]. Japkowicz et al. [35] used sampling approaches to perform tests and then analyzed data using different categorization techniques to assess the impact of imbalanced data [37].

These investigations demonstrated the significance of conceptual challenges and demonstrated that a domain's categorization could not be influenced by a large CI provided the idea was simple to training. Therefore, the goal of the resolution needs to be to grasp two crucial criteria - degree of CI and level of difficulty of the idea.

2. Literature Review

Wang, M., Yao, X., & Chen, Y. (2021), the undersampling clustering-oversampling method (shortly, UCO algorithm) is made by bringing together random undersampling, clustering, and oversampling approaches [38]. The UCO technique gives near about balanced data that is utilized for training machine-learning models for assessing heart attack probability. The random forest classifier shown excellent results of prediction, with accuracy of 70.29% and a precision of 70.05%. One can

find out if a stroke patient will have a heart attack or not using UCO (120) and random forest. [1].

Zhang, S., Fan, L., Ye, J., Chen, G., Fu, C., & Leng, Y. (2023) narrates the 6-minute walking the distance travelled and the Lower-limb Fugl-Meyer assessment in stroke patients gives key standards for finding the effectiveness of instruction and guiding training initiatives. At last, the assessment model was guided using a machine-learning method. The trained model finds the 6-minute walking distance and Lower-limb Fugl-Meyer values with prediction accuracy of 85.19% and 92.66%, respectively. Values of fresh stroke patients. From experiments, it was found that the assessment accuracy for the 6MWD and FMA-LE values are 85.19% and 92.66%, respectively [40]

Lin, P., Zhai, X., Li, W., Li, T., Cheng, D., Li, C., Pan, Y., & Ji, L. (2022) utilized a deep learning forecast model recognized for another treatment to assess the chance of expecting the result of a treatment [16]. This study was conducted on 15 stroke survivors, who have their clinical and physiological data checked before and after therapy (clinical measurement, biomechanical measurement, and electroencephalography (EEG) measurement). Many biomarkers and clinical scale scores were checked in persons who completed manual stretching rehabilitation treatment. Data were used to construct deep learning prognosis models that shown a prognosis accuracy of 87.50% [16].

Fang, G., Xu, P., & Liu, W. (2020), on The International Stroke Trial (IST) dataset, and combined machine learning technique applied to classify the subtype of ischemic stroke. In this paper, they checked at the often difficulties of feature selection and prediction in medical datasets. Initially, the Shapiro-Wilk method was utilized to categorize the relevance of characteristics, and Pearson correlations between features were examined. Then, they utilize Recursive Feature Elimination with Cross-Validation (RFECV) to predict robust features important for ischemic stroke subtyping that included, Random-Forest-Classifier, Extra-Trees-Classifier, linear SVC, AdaBoost-Classifier, and Multinomial-Nave-Bayes-Classifier as predictors. [14].

Indarto, I., Utami, E., & Raharjo, S. (2020) inferred Hemorrhagic stroke has high death rate than ischemic stroke. The aim of this work is to make a mortality prediction model relied on demographic and clinical data in patients having hemorrhagic stroke utilizing Multilayer Perceptron algorithms, J48 and Jrip, . The

forecast of stroke mortality helps physicians in finding prognosis, targeted therapy, and preparing patients and families. Apart from ischemic stroke patients, 1329 participants were grouped from Yogyakarta's Hospital Stroke Registry [15].

Pathan, M. S., Zhang, J., John, D., Nag, A., & Dev, S. (2020) recommended approach determine important stroke signs from large dataset for creating outstanding stroke prediction models. Additionally, it changed the rough sets approach in order to be used on medical datasets with binary feature sets relating to real-world applications that were earlier not used with the old rough sets technique. It utilized the recommended approach on a dataset of 29, 072 patient records with 10 shared features. It discovered four most crucial features, that is age, heart disease, hypertension, and average glucose level that forms the most to find out stroke diagnosis in comparison to rest attributes [7].

Lee, W., & Seo, K. (2022) stated that for a binary classification problem with skewed classes, the active learning technique is utilized for getting better performance with a high balanced class distribution [10]. The ongoing downsampling strategy is used to reduce the generalization error of model forecasting and consecutive model upgrading by picking good instances one at a time. It focused a number of subjects for further work. First, in the last stated experiment, the number of samples k for downsampling is informed in advance. Though, categorization performance is better by a point before plateauing. It is good to find out the point at which performance starts to congregate without the need for additional instance selection. Also logistic regression works good for binary classification issues in general but is constrained to linearity. It expand the recommended active downsampling model above linearity by using more agile models, such as smoothing spline. Third, while the suggested technique emphasizes on practice, binary imbalance categorizing is a common classifying difficulty, based on this work, this method may be applied to multi-class imbalance classification. It concentrates on classification with a binary imbalance [10].

Hoyos-Osorio, J., Álvarez-Meza, A. M., Daza-Santacoloma, G., Orozco-Gutierrez, Á., & Castellanos- Domínguez, G., (2021) proposed Relevant Information-based UnderSampling (RIUS) method for choosing the related samples from the majority class to improve classification performance in unbalanced data settings. RIUS is founded on the information-preservation concept that takes the below structure of the majority class utilizing less samples. Additionally, it combine

RIUS technique with the renowned Clustering-based Undersampling algorithm (CBUS) for better data representation, and called it RIUS upgrade CRIUS. The test results recommends that RIUS and CRIUS reveal the essential structure of the data and stops information loss by choosing the high informative occurrences. It has a lacuna that it works well for binary class jobs [38].

Mathew, J., Pang, C. K., Luo, M., & Leong, W. H. (2018) gives a weighted kernel-relied SMOTE (WK-SMOTE) which by oversampling in the feature space of a support vector machine (SVM) classifier, takes care of the restriction of SMOTE for nonlinear situations. On many of benchmark unbalanced data sets, the recommended oversampling approach, with a cost-sensitive SVM formulation, gives better results than baseline methods. SVM classifier merely balances the class distribution [42].

I. Almomani et al., (2021) suggests alternate for finding Ransomware which rests on an evolutionary-based machine learning approach. The binary particle swarm optimization approach is utilized to adjust the classification algorithm's hyperparameters and to perform feature selection. For segregation, the support vector machines (SVM) method is utilized with the synthetic minority oversampling technique (SMOTE) [1]. The used dataset is a combination of 10,153 Android apps, 500 of which are Ransomware. The impact of the recommended method SMOTE-tBPSO-SVM gave better results than usual machine learning methods with high sensitivity, specificity, and g-mean scores. It does not rests on more data or better models to treat huge data [41] [43].

3. Objectives of Research

The objectives of this research for developing a thesis on "Stroke Prediction Using Imbalanced Classification" outlines the specific goals and outcomes to be achieved from the research work. These objectives will guide research process and help to stay focused on the key aspects of thesis. Here are some objectives that are planned for research work:

- Develop a predictive model that can accurately classify individuals into stroke risk categories, with a focus on improving sensitivity and specificity for the minority class in the imbalanced dataset.
- Implement and evaluate technique for handling class imbalance to mitigate the impact of imbalanced data on model performance.
- Identify and select the most relevant features or risk factors associated with stroke, using feature selection and engineering techniques to enhance the model's predictive power.
- Experiment with machine learning algorithms to determine the best predictive performance.
- Assess the interpretability of the developed model by analyzing feature importance scores and employing techniques provide insights into prediction rationale.
- Investigate potential biases in model predictions and explore strategies to ensure fairness, especially concerning protected attributes (e.g., gender).
- Address ethical considerations related to healthcare data privacy, informed consent, and responsible AI usage.
- Validate the model's performance on an independent dataset to assess its ability to generalize to new data.

4. Methodology

Developing the methodology for the thesis on "Stroke Prediction Using Imbalanced Classification" is a crucial step in the research process. Here's a comprehensive Methodology for Stroke Prediction Using Imbalanced Classification for the thesis:

4.1 Data Collection

- Data Sources: Identify and collect relevant medical datasets that include information on patient demographics, medical history, lifestyle factors, and stroke outcomes. Ensure the datasets encompass a diverse population.
- Data Preprocessing: Clean the data by addressing missing values, outliers, and data quality issues. Verify data integrity and ensure it is suitable for analysis.

- Data Split: Divide the dataset into training, validation, and test sets, ensuring the class imbalance is preserved in each subset.

4.2 Imbalanced Data Handling

To solve the class imbalance issue, explore different unbalanced classification strategies such as oversampling [22] [23] [24], undersampling [19] and ensemble methods. Implement and tailor these strategies to the unique needs of stroke prediction.

4.3 Feature Selection and Engineering

Feature Selection: Identify relevant features using techniques like Recursive Feature Elimination (RFE) or feature importance analysis from tree-based models. Feature Engineering: Create new features or transformations that may enhance the predictive power of the model, such as age categories or risk factor scores.

4.4 Model Development

Using cutting-edge methods such as logistic regression, decision trees, random forests, or support vector machines [23], create an unbalanced classification model [3]. To improve performance, fine-tune the model parameters and feature selection techniques [5].

4.5 Evaluation Metrics

Use suitable evaluation metrics to assess the constructed stroke prediction model, such as accuracy, precision, recall, F1-score, and AUC-ROC [27]. To evaluate the efficacy and superiority of the suggested method, compare the results to those of current approaches.

4.6 Integration, Validation and Testing

Integrate the model into clinical workflows, ensuring it can provide real-time risk assessment. Validate the model's performance on an independent dataset to assess its generalizability. Conduct rigorous testing and validation to ensure the model's reliability and safety for clinical use.

4.7 Proposed work plan

Table 1: Work plan for the year and goals to attain

Year	Work	Research Objectives
I (2023-24)	**Literature Review** • Review existing literature on stroke prediction, classification techniques, and imbalanced datasets. • Identify key studies, datasets, and methodologies related to stroke prediction. • Analyze challenges in stroke prediction, including class imbalance, and existing solutions. • Highlight gaps in current research and explain how your thesis addresses them. **Data Collection and Preprocessing** • Gather relevant medical data and stroke-related datasets. • Clean and preprocess the data to handle missing values, outliers, and ensure data quality. • Address class imbalance techniques.	RO1
II (2023-24)	**Feature Selection and Engineering** • Identify important features related to stroke risk. • Perform feature selection to reduce dimensionality. • Engineer new features or transformations to enhance predictive power. **Model Development** • Experiment with various machine learning algorithms for stroke prediction. • Evaluate the performance of each algorithm on imbalanced data [39].	RO2

	• Implement advanced ensemble techniques to combine models for improved accuracy. **Evaluation Metrics and Model Selection** • Select appropriate evaluation metrics for imbalanced classification (e.g., precision, recall, F1-score, AUC-ROC). • Study the trade-off between sensitivity and specificity. • Compare and analyze the results of different models and techniques [37]. **Validation and Testing** • Validate the model's performance on an independent dataset. • Conduct rigorous testing to ensure the model's reliability and safety. • Document the validation results and any necessary adjustments to the model.	
III (2024-25)	• Pre Ph.D. Synopsis Submission • Thesis writing, Thesis Submission	RO3

5. Ethical Considerations

Ensure compliance with ethical guidelines, healthcare regulations, and data protection laws throughout the research process. Ensure that data used in the study complies with ethical guidelines and data privacy regulations.

References

1. Wang, M., Yao, X., & Chen, Y., "An Imbalanced-Data Processing algorithm for the prediction of heart attack in stroke patients", IEEE Access, 9, 25394–25404, 2021, https://doi.org/10.1109/access.2021.3057693
2. Dev, S., Sun, L., Nwosu, C. S., Jain, N., Veeravalli, B., &Deepu, C. J., " A predictive analytics approach for stroke prediction using machine learning and neural networks", Healthcare Analytics, 2, 100032, 2022, https://doi.org/10.1016/j.health.2022.100032
3. Kashi, S., Feingold-Polak, R., Lerner, B., Rokach, L., & Levy-Tzedek, S., " A Machine-Learning model for Automatic Detection of movement compensations in stroke patients", IEEE Transactions on Emerging Topics in Computing, 9(3), 1234–1247, 2021, https://doi.org/10.1109/tetc.2020.2988945
4. Ivanov IG, Kumchev Y, Hooper VJ., "An Optimization Precise Model of Stroke Data to Improve Stroke Prediction. Algorithms", 2023; 16(9):417. https://doi.org/10.3390/a16090417
5. Newaz, M. A., Hassan, M. S., &Haq, F., "An empirical analysis of the efficacy of different sampling techniques for imbalanced classification", arXiv (Cornell University), 2022, https://doi.org/10.48550/arxiv.2208.11852
6. Bosch, K. (n.d.). Secondary Stroke Prevention and Support: an occupational therapy program. UND Scholarly Commons. https://commons.und.edu/ot-grad/377/
7. Pathan, M. S., Zhang, J., John, D., Nag, A., & Dev, S., "Identifying stroke indicators using rough sets", IEEE Access, 8, 210318–210327, 2020 https://doi.org/10.1109/access.2020.3039439
8. Hwang, S. Y., Jo, I. J., Lee, S. U., Lee, T. R., Yoon, H., Cha, W. C., Sim, M. S., & Shin, T. G., "Low accuracy of positive QSOFA criteria for predicting 28-Day mortality in critically ill septic patients during the early period after emergency department presentation", Annals of Emergency Medicine, 71(1), 1-9.e2, 2018, https://doi.org/10.1016/j.annemergmed.2017.05.022
9. Mail, D. (2005, April 12). Scan predicts heart attacks in advance. Mail Online. https://www.dailymail.co.uk/health/article-344598/Scan-predicts-heart-attacks-advance.html

10. Lee, W., & Seo, K., “Downsampling for Binary Classification with a Highly Imbalanced Dataset Using Active Learning”, Big Data Research, 28, 100314, 2022, https://doi.org/10.1016/j.bdr.2022.100314
11. K, N. A., Anisha, C., K, A. S., Shaji, J., & Kamalini, T., “Design and implementation of an automatic physiotherapy device for ankle”, International Journal for Research in Applied Science and Engineering Technology”, 11(5), 417–419, (2023), https://doi.org/10.22214/ijraset.2023.51417
12. Hui, H., Zhang, X., Li, F., Mei, X., & Guo, Y., “A Partitioning-Stacking prediction fusion network based on an improved attention U-Net for stroke lesion segmentation”, IEEE Access, 8, 47419–47432, 2020, https://doi.org/10.1109/access.2020.2977946
13. Shubham Avadhesh Pandey , Manjit Kumar, “A case study on post-stroke physiotherapy management”, International Journal of Novel Research and Development (www.ijnrd.org), ISSN:2456-4184, Vol.8, Issue 5, page no.f725-f741, May-2023, Available :http://www.ijnrd.org/papers/IJNRD2305594.pdf
14. Fang, G., Xu, P., & Liu, W., “Automated Ischemic Stroke Subtyping based on Machine learning approach”, IEEE Access, 8, 118426–118432, 2020 https://doi.org/10.1109/access.2020.3004977
15. Indarto, I., Utami, E., & Raharjo, S., “Mortality prediction using data mining classification techniques in patients with hemorrhagic stroke”, 8th International Conference on Cyber and IT Service Management (CITSM), 2020, https://doi.org/10.1109/citsm50537.2020.9268802
16. Lin, P., Zhai, X., Li, W., Li, T., Cheng, D., Li, C., Pan, Y., & Ji, L., “A transferable deep learning prognosis model for predicting stroke patients’ recovery in different rehabilitation trainings”, IEEE Journal of Biomedical and Health Informatics, 26(12), 6003–6011, 2022, https://doi.org/10.1109/jbhi.2022.3205436
17. Apollo Hospitals Navi Mumbai. (2021, July 1). Stroke treatment in Navi Mumbai - Apollo Hospitals Mumbai. Apollo Hospital Mumbai. https://www.apollohospitals.com/mumbai/clinical-excellence/neurosciences/treatments-and-procedures/neurosurgery/stroke/
18. Staying Stroke Smart: Promoting awareness and Prevention. (n.d.). Feast. https://vocal.media/feast/staying-stroke-smart-promoting-awareness-and-prevention-90580uw7

19. Wang, X., Liu, B., Cao, S., Jing, L., & Yu, J. Z., "Important sampling based active learning for imbalance classification", Science China Information Sciences, 63(8), SringerLink, 2020, https://doi.org/10.1007/s11432-019-2771-0
20. H. He, and E. Garcia, "Learning from imbalanced data", IEEE Transactions on Knowledge & Data Engineering, vol. 9, pp. 1263-1284, 2008.
21. C. Pannucci and E. Wilkins, "Identifying and avoiding bias in research. Plastic and reconstructive surgery", vol. 126, p.619, 2010.
22. N.V. Chawla, "Data mining for imbalanced datasets: An overview", In Data mining and knowledge discovery handbook, pp. 875-886, 2009.
23. F. Alberto, S. García, M. Galar, R. Prati, B. Krawczyk, and F. Herrera, "Learning from imbalanced data sets", Springer, 2018.
24. Luque, A., Carrasco, A., Grosicki, S., & De Las Heras, A. I., "The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition", 91, 216–231, Elsevier, 2019, https://doi.org/10.1016/j.patcog.2019.02.023
25. Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M., "Boosting methods for multi-class imbalanced data classification: an experimental review", Journal of Big Data, 7(1) SpringerOpen, 2020, https://doi.org/10.1186/s40537-020-00349-y
26. M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection", In Proceedings of the Fourteenth International Conference on Machine Learning, vol. 97, pp. 179-186, 1997.
27. Yao, P., Shen, S., Xu, M., Liu, P., Zhang, F., Xing, J., Shao, P., Kaffenberger, B. H., & Xu, R. X., "Single Model Deep Learning on Imbalanced Small Datasets for Skin Lesion Classification", IEEE Transactions on Medical Imaging, 41(5), 1242–1254, 2022, https://doi.org/10.1109/tmi.2021.3136682
28. P. Mrozek, J. Panneerselvam and O. Bagdasar. Efficient Resampling for Fraud Detection during Anonymised Credit Card Transactions with Unbalanced Datasets. 2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC), Leicester, UK, 2020, pp. 426-433, doi: 10.1109/UCC48980.2020.00067.

29. Padurariu, C., & Breaban, M., "Dealing with Data Imbalance in Text Classification", Procedia Computer Science, 159, 736–745, Elsevier, 2019, https://doi.org/10.1016/j.procs.2019.09.229
30. P. S. Thakur, M. Jadeja and S. S. Chouhan, "Text Augmentation based Imbalance Learning for Unstructured Text Data", IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), Goa, India, pp. 73-77, 2022, doi: 10.1109/ICCCMLA56841.2022.9989047. https://ieeexplore.ieee.org/abstract/document/9989047
31. Parlak, B., "Class index corpus index measure: A novel feature selection method for imbalanced text data. Concurrency and Computation: Practice and Experience", 34(21), Wiley Online Library 2022, https://doi.org/10.1002/cpe.7140
32. Chaubey, G., Gavhane, P. R., Bisen, D., & Arjaria, S. K., "Customer purchasing behavior prediction using machine learning classification techniques", Journal of Ambient Intelligence and Humanized Computing, 2022, https://doi.org/10.1007/s12652-022-03837-6
33. Chaabane, I., Guermazi, R., & Hammami, M., "Enhancing techniques for learning decision trees from imbalanced data", Advances in Data Analysis and Classification, 14(3), 677–745, SpringerLink, 2020, https://doi.org/10.1007/s11634-019-00354-x
34. Kumar, P., Bhatnagar, R., Gaur, K., & Bhatnagar, A., "Classification of Imbalanced Data:Review of Methods and Applications", IOP Conference Series: Materials Science and Engineering, 1099(1), 012077, 2021. https://doi.org/10.1088/1757-899x/1099/1/012077
35. N.Japkowicz and S.Stephen. The class imbalance problem: A systematic study. Intelligent Data Analysis, vol. 6, pp.429-449, 2002.
36. Baldé, M. A., Balde, C., & Ndiaye, B. M., "Impact studies of nationwide measures COVID-19 anti-pandemic: compartmental model and machine learning", Communications in Mathematical Biology and Neuroscience, 2020. https://doi.org/10.28919/cmbn/4911
37. Xu, Z., Shen, D., Nie, T., & Kou, Y., "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data",

Journal of Biomedical Informatics, 107, 2020, https://doi.org/10.1016/j.jbi.2020.103465

38. Hoyos-Osorio, J., Álvarez-Meza, A. M., Daza-Santacoloma, G., Orozco-Gutierrez, Á., & Castellanos- Domínguez, G., "Relevant information undersampling to support imbalanced data classification", Elsevier, Neurocomputing, Volume 436, Pages 136- 146, 14 May 2021, https://doi.org/10.1016/j.neucom.2021.01.033
39. Li, X., Wang, L., & Sung, E., "AdaBoost with SVM-based component classifiers" Engineering Applications of Artificial Intelligence, 21(5), 785–795, 2008, https://doi.org/10.1016/j.engappai.2007.07.001
40. Zhang, S., Fan, L., Ye, J., Chen, G., Fu, C., & Leng, Y., "An Intelligent Rehabilitation Assessment Method for Stroke Patients based on Lower Limb Exoskeleton Robot", IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2023
41. Rawat, S. S., & Mishra, A. K., "Review of Methods for Handling Class-Imbalanced in Classification Problems", arXiv (Cornell University). https://doi.org/10.48550/arxiv.2211.05456
42. Mathew, J., Pang, C. K., Luo, M., & Leong, W. H., "Classification of imbalanced data by oversampling in kernel space of support vector machines" IEEE Transactions on Neural Networks and Learning Systems, 29(9), 4065–4076, 2018, https://doi.org/10.1109/tnnls.2017.2751612
43. I. Almomani et al., "Android Ransomware Detection Based on a Hybrid Evolutionary Approach in the Context of Highly Imbalanced Data", in IEEE Access, vol. 9, pp. 57674-57691, 2021, doi: 10.1109/ACCESS.2021.3071450.

Ms. Manisha Amrutrao Manjramkar — Research Scholar

Dr. Kalpana Jondhale_Paithane — Research Supervisor